VISUAL AND COMPUTATIONAL CONSIDERATIONS IN SMOOTHING SCATTERPLOTS BY ROBUST LOCALLY WEIGHTED REGRESSION

William S. Cleveland

Bell Laboratories
Murray Hill, New Jersey 07974

ABSTRACT

The scatterplot is one of the most powerful and most used statistical tools. With only a small amount of additional effort, the visual information can be greatly increased by plotting another set of points whose purpose is to summarize some aspect of the scatterplot.

1. INTRODUCTION

Figure 1 shows a scatterplot of points (x_i,y_i), for $i = 1,...,n$, where $n = 50$. In Figure 2 the same scatterplot is summarized by another set of points $(x_i,\hat{y}_i)$, for $i = 1,...,n$, which are plotted by joining successive values by straight lines. The point $(x_i,\hat{y}_i)$ portrays the middle of the distribution of the variable on the vertical axis, Y, given the value of the variable on the horizontal axis, $X = x_i$. The formation of the new points will be referred to as "smoothing" the scatterplot. The point $(x_i,\hat{y}_i)$ is called the smooth at x_i and $\hat{y}_i$ is called the fitted value at x_i.

The example in Figure 1 was generated by taking $x_i = i$, for $i = 1,..., 50$ and

$$y_i = .02\ x_i + \epsilon_i\ ,$$

where the ϵ_i are a random sample from a normal distribution with mean 0 and variance 1. The linear effect is not easily perceived from the scatterplot alone, but is revealed when the smooth is superimposed.

In this paper we shall discuss a method for smoothing scatterplots called robust locally weighted regression. The details of the method are given in Section 2. Various visual considerations and alternative plotting procedures are discussed in Section 3 and computational matters are discussed in Section 4. References [1], [2], [3, p. 225], [4], [5, Chapters 8 and 9], and [6] describe other methods for smoothing scatterplots.

2. ROBUST LOCALLY WEIGHTED REGRESSION

The method of smoothing used in Figure 2, which is called robust locally weighted regression, is defined by the following sequence:

(1) Let

$$W(x) = (1-|x|^3)^3 I(x)$$

where $I(x) = 1$ if $|x| \leqslant 1$ and $I(x) = 0$ if $|x| > 1$. Let

$$B(x) = (1-x^2)^2 I(x)\ .$$

(2) For each i let h_i be the distance from x_i to the r-th nearest neighbor of x_i. That is h_i is the r-th smallest number among $|x_i-x_j|$, for $j = 1,...,n$. For $k = 1,...,n$ let

$$w_k(x_i) = W\left[h_i^{-1}(x_k-x_i)\right]\ .$$

(3) For each i compute $\hat{\beta}_0(x_i)$ and $\hat{\beta}_1(x_i)$, the intercept and slope respectively, of a linear regression of y_k on x_k using weighted least squares with weight $w_k(x_i)$ at (x_k,y_k). That is, $\hat{\beta}_0(x_i)$ and $\hat{\beta}_1(x_i)$ are the values of β_0 and β_1 which minimize

$$\sum_{k=1}^{n} w_k(x_i)(y_k-\beta_0-\beta_1 x_k)^2\ .$$

Let

$$\hat{y}_i = \hat{\beta}_0(x_i) + \hat{\beta}_1(x_i)x_i$$

be the fitted value of the line at x_i.

(4) Let

$$e_i = y_i - \hat{y}_i$$

be the residuals from the current fitted values. Let s be the median of the $|e_i|$. Define robustness weights by

$$\delta_k = B\left(\frac{e_i}{6s}\right)\ .$$

(5) Recompute $\hat{y}_i$ for each i by fitting a line using weighted least squares with weight $\delta_k w_k(x_i)$ at (x_k,y_k).

(6) Repeatedly carry out steps (4) and (5) a total of one, two, or three times or until convergence occurs. The

final $\hat{y}_i$ are robust locally weighted regression fitted values.

The weights $w_k(x_i)$ decrease as the distance of x_k from x_i increases. Thus points whose abscissas are close to x_i play a large role in the determination of $\hat{y}_i$ while points far away play little or no role. Increasing r, the number of nearest neighbors, tends to increase the smoothness of the smoothed points $(x_i,\hat{y}_i)$. Choosing r to be 20 to 80 of n should serve most purposes. A practical default value, used in the example of Figures 1 and 2 is $.5n$. The "tricube" weight function, $W(x)$, is used to weight neighbors since it results in certain desirable statistical properties [7].

The iterative fitting in steps (4) to (6) is carried out to achieve a robust smooth in which a small fraction of deviant points does not distort the results. Deviant points tend to have small robustness weights, δ_k, and therefore do not play a large role in the determination of the smoothed values. The "bisquare" weight function, $B(x)$, is used since other investigations have shown it to perform well for robust estimation of location [8] and for robust regression [9]. Two iterations of steps (4) and (5) are generally quite sufficient; this is the number used in the example of Figures 1 and 2.

3. VISUAL CONSIDERATIONS

3.1 Plotting the Smooth

The smoothed points can be plotted by joining successive points by straight lines as in Figure 2 or by symbols at the points $(x_i,\hat{y}_i)$. When the smooth is superimposed on the scatterplot the first method provides greater visual discrimination with the points of the scatterplot. But using lines raises the danger of an inappropriate interpolation. One possible approach is to use symbols initially when analyzing the data; then if a particular plot is needed for further use, such as presentation to others, the lines can be used if the initial plot indicates that linear interpolation would not lead to a distortion of the results.

The smoothed values also can be plotted on a separate grid with the same scales as the original scatterplot. This is particularly attractive for low resolution plots such as printer plots.

3.2 Symmetric Summaries

The method of summarizing the scatterplot in Section 2 is appropriate when Y is the response or dependent variable and X is the independent variable. In cases where neither variable can be designated as the response, the scatterplot can be summarized by plotting the smooth of Y given X and the smooth of X given Y.

3.3 Summarizing Scale and Choosing a Scale Stabilizing Transformation

The smoothed points in Figure 1 portray the location of the distribution of Y given $X = x_i$. It is often useful to have, in addition, a summary of the scale. This can be done by plotting $|y_i - \hat{y}_i|$ against x_i and computing and plotting a smooth, $(x_i,\hat{s}_i)$, of this scatterplot.

If the scale of y_i is a function, $\sigma(\mu)$, of the location of y_i then the transformation of y_i which stabilizes the scale [10, p. 425] is

$$t = \int \frac{1}{\sigma} .$$

Suppose t is a power transformation

$$t(\mu) = \begin{cases} \dfrac{\mu^p - 1}{p} & p \neq 0 \\ \log \mu & p = 0 \end{cases} .$$

Tukey [5, p. 103] has suggested a procedure for choosing a scale stabilizing power transformation for batches of numbers, which can be extended to choosing one for scatterplots. From the above equations we have

$$\log \sigma(\mu) = -\log t'(\mu) = -(p-1)\log \mu .$$

A plot of $\hat{s}_i$ vs. $\hat{y}_i$ describes the function $\sigma(\mu)$. Thus a plot of $\log \hat{s}_i$ vs. $-\log \hat{y}_i$ will, apart from sampling fluctuations, follow a line with slope $p-1$. Thus p can be chosen by fitting a line to the plot, either by eye or by some numerical method.

3.4 Judging the Amount of Smoothing

The most practical method for choosing r, the number of nearest neighbors, is to study the visual display. The objective is to choose r as large as possible without distorting the underlying pattern in the scatterplot.

The fitted value, $\hat{y}_i$, in step (3) of Section 2 can be written as

$$\hat{y}_i = \sum_{k=1}^{n} r_k(x_i) y_k ,$$

where $r_k(x_i)$ depends only on $x_1, \ldots, x_n$. The equivalent number of parameters

$$enp = 2 \sum_{i=1}^{n} r_i(x_i) - \sum_{i,k=1}^{n} r_k^2(x_i)$$

also can be used to judge the relative amounts of smoothing for different values of r. An interpretation of enp arises from considering the variability in the residuals, e_i, of the fitted values in (3). Suppose the y_i are independent and have common variance σ^2 then

$$enp = n - \frac{E \sum_{i=1}^{n} e_i^2}{\sigma^2} .$$

Suppose the e_i were the residuals from a linear least squares regression of y_i on x_i using q parameters, then enp would be equal to q. Thus enp for locally weighted regression can be interpreted as an equivalent number of parameters.

In [7] it is shown that enp can be approximated (when the weight function is tricube) by

$$2(1 + \frac{n}{r}) .$$

Thus for the default value of $\frac{r}{n} = .5$ described in Section 2, the approximate equivalent number of parameters is 6.

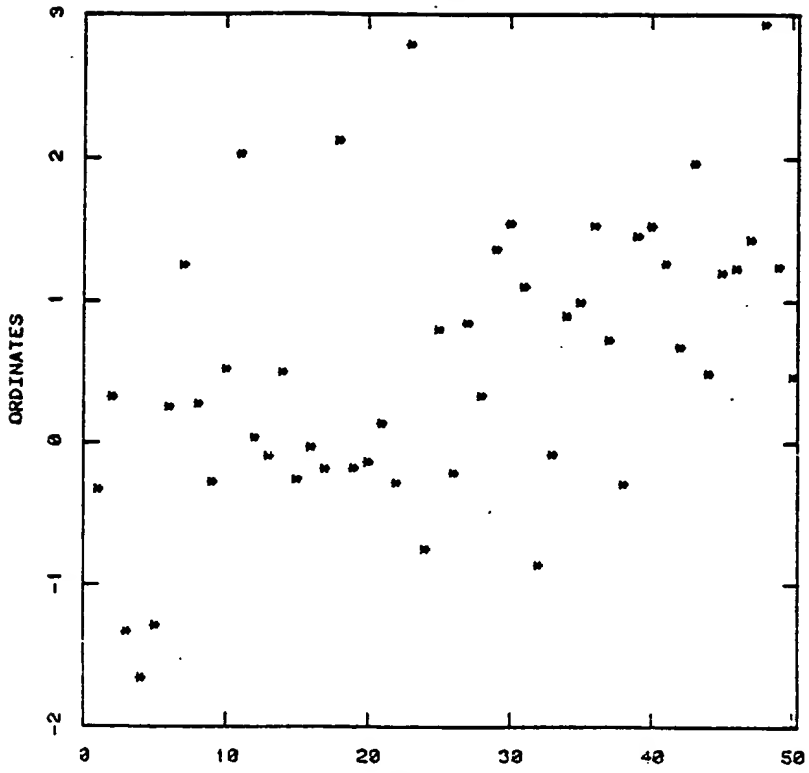

FIGURE 1

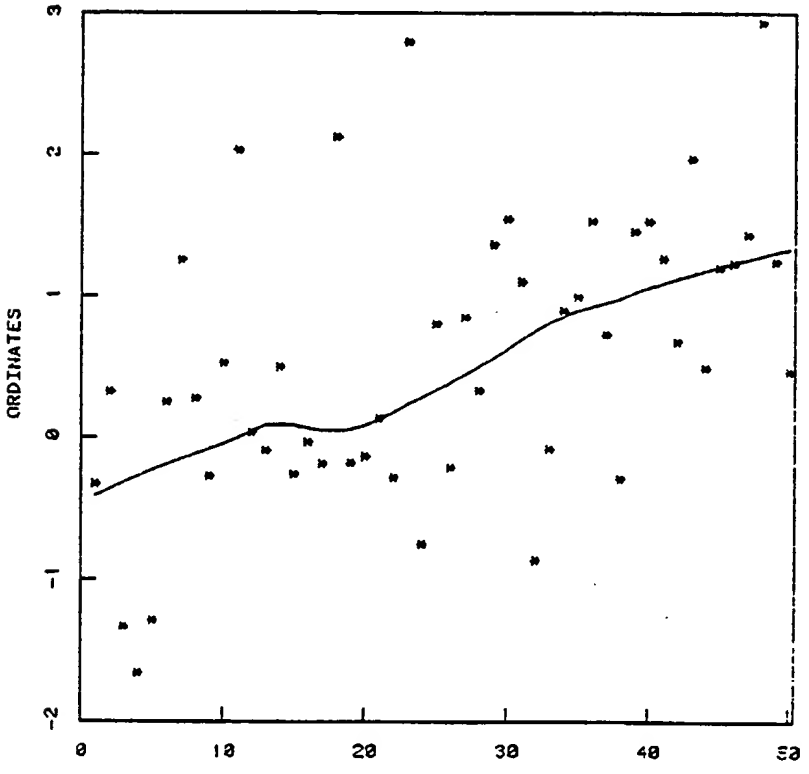

FIGURE 2

3.5 Nearest Neighbor vs. Equal Resolution

An alternative to choosing h_i through nearest neighbors is to use a constant value h for the computation of all fitted values. This provides equal resolution over all regions of the scatterplot but leads to appreciable increases in the variance at isolated points and at the ends of the scatterplot. The selection of the nearest neighbor routine is based on its more satisfactory performance, particularly at the ends of the scatterplot, for the applications which I have encountered. However, other users may find equal resolution more satisfactory in other applications.

4. COMPUTATIONAL CONSIDERATIONS

4.1 Reducing the Computations

Suppose the x_i are ordered from smallest to largest and let $x_{a(i)}, \ldots, x_{b(i)}$ be the ordered r nearest neighbors of x_i. The values of $a(i+1)$ and $b(i+1)$ can be found from $a(i)$ and $b(i)$ using the following scheme:

(1) Let $A = a(i)$ and $B = b(i)$.

(2) Let $d_A = x_{i+1} - x_A$ and $d_B = x_{B+1} - x_{i+1}$.

(4) a. If $d_A \leqslant d_B$ then $a(i+1) = A$ and $b(i+1) = B$.
b. If $d_A > d_B$ replace A by $A+1$ and B by $B+1$ and return to (2).

(4) h_{i+1} is the maximum of $x_{i+1} - x_A$ and $x_B - x_{i+1}$.

Thus this scheme can be used to save computations by computing the fitted values at x_1, then x_2, etc. Only $x_{a(i)}, \ldots, x_{b(i)}$ need be considered in the weighted least squares computation of $\hat{y}_i$ since $W(x) = 0$ for $|x| \geqslant 1$. This saving would not be achieved by using a weight function which becomes small but not zero for large x, such as the normal probability density.

4.2 Computation Time

An experiment was run to determine the run time of the smooth using the scheme in Section 4.1 for the nearest neighbor algorithm in Section 2 with one iteration. (The portable FORTRAN routine is available from the author.) Each additional iteration would increase the time by slightly less than 50 of the time required for one iteration. Consideration of the algorithm shows that the run time is independent of the configuration of the x_i. (This would not be true for the equal resolution algorithm.) The experiment was run for all 20 combinations of 5 values of $f = r/n$ (.2, .4, .6, .8, 1.0) and 4 values of n (25, 50, 100, and 200). A least squares fit to the log (base 10) run time (cpu milleseconds) resulted in the fitted equation

$$\log \textit{time} = -.49 + 1.98 \log n + .87 \log f .$$

The estimate of the residual standard error is the standard error of the log times is .718 log milleseconds. Thus the equation provides a very close fit to the data.

4.3 Thinning

The computations for the nearest neighbor algorithm are, as shown in the previous section, approximately of the order $f^{.9} n^2$. For scatterplots with fewer than 50 to 100 points the computations present no problems. Plots with more points generally need not incur the cost of using all the points since computing the smooth at a subset of the points will generally perform satisfactorily. (The smooth can, of course, still be superimposed on the full scatterplot.)

Two possible methods of thinning are to select every t-th value of the ordered x_i or to form a grid of equally spaced points on the horizontal axis and select, for each grid value, the x_i which is closest to the value.

4.4 Locally Weighted Regression of Order d

Steps (3) and (5) of the procedure in Section 2 can be generalized by fitting a polynomial of degree d, where d is a non-negative integer. Choosing $d = 1$ appears to strike a good balance between computational ease and the need for flexibility to reproduce patterns in the data. The case $d = 0$ is the simplest, computationally, but in the practical situation an assumption of local linearity seems to serve far better than an assumption of local constancy since the practice is to plot variables which are related to one another. For $d = 2$, however, computational considerations begin to override the need for having flexibility.

ACKNOWLEDGEMENT

The computing considerations for smoothing scatterplots were, in part, shaped by some very helpful discussions with my colleague, Rick Becker.

William S. Cleveland

REFERENCES

[1] Clark, R. M. (1977). "Non-parametric Estimation of a Smooth Regression Function," *Journal of the Royal Statistical Society*, Series B, 39, 107-113.

[2] Cleveland, W. S., and Kleiner, B. (1975). "A Graphical Technique for Enhancing Scatterplots with Moving Statistics," *Technometrics*, 17, 447-454.

[3] Ezekiel, M. (1941). *Methods of Correlation Analysis*, Second Edition, Wiley, New York.

[4] Stone, C. J. (1977). "Consistent Nonparametric Regression," *The Annals of Statistics*, 5, 595-620.

[5] Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading, Mass.

[6] Watson, G. S. (1964). "Smooth Regression Analysis," *Sankhya*, 26, Series A, 359-372.

[7] Cleveland, William S. (1977). "Locally Weighted Regression and Smoothing Scatterplots," (submitted for publication).

[8] Gross, A. M. (1976). "Confidence Interval Robustness with Long-Tailed Symmetric Distributions," *Journal of the American Statistical Association*, 71, 409-416.

[9] Gross, A. M. (1977). "Confidence Intervals for Bisquare Regression Estimates," *Journal of the American Statistical Association*, 72, 341-354.

[10] Kendall, M. G. and Stuart, A. S. (1977). *The Advanced Theory of Statistics*, Volume 1, 4th Edition. Hafner, New York.